Simplifying Generative AI App Development: Why Standards Matter

Katherine Druckman

Open Source Security Evangelist

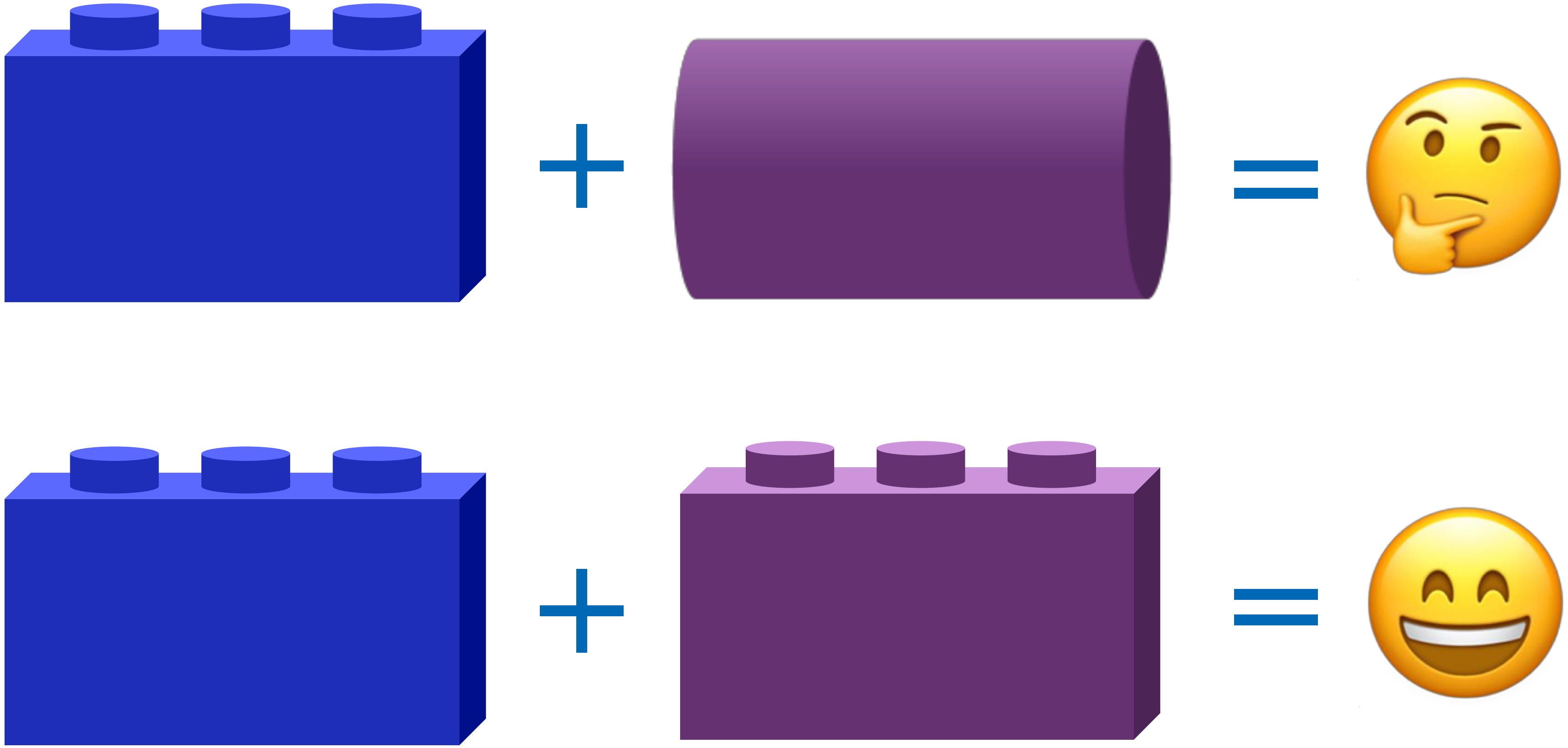
Ezequiel Lanza

Open Source AI Evangelist

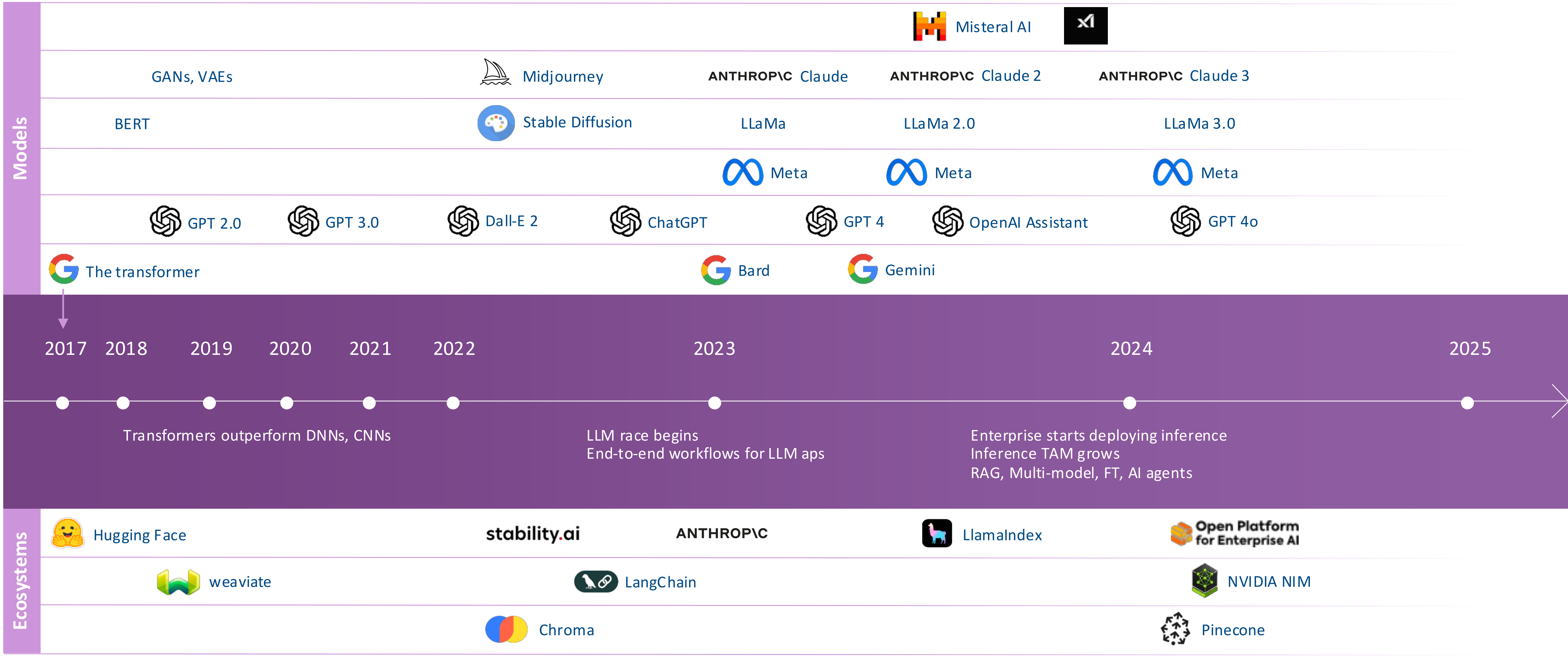
GenAI is emerging rapidly, but enterprises are struggling to realize GenAI value in production.

Why standardization matters

Life is easier when things go together



GenAI: transformers to AI Agents in 8 years



Models

Misteral AI

GANs, VAEs

Midjourney

ANTHROP\C Claude

ANTHROP\C Claude 2

ANTHROP\C Claude 3

BERT

Stable Diffusion

LLaMa

LLaMa 2.0

LLaMa 3.0

Meta

Meta

Meta

GPT 2.0

GPT 3.0

Dall-E 2

ChatGPT

GPT 4

OpenAI Assistant

GPT 4o

The transformer

Bard

Gemini

2017 2018 2019 2020 2021 2022 2023 2024 2025

Transformers outperform DNNs, CNNs

LLM race begins
End-to-end workflows for LLM aps

Enterprise starts deploying inference
Inference TAM grows
RAG, Multi-model, FT, AI agents

Ecosystems

Hugging Face

stability.ai

ANTHROP\C

LlamaIndex

Open Platform for Enterprise AI

weaviate

LangChain

NVIDIA NIM

Chroma

Pinecone

The price of unbridled exploration

Massive innovation, expensive duplication, no standardization

- So many options, how do decide?
- Constantly reinventing the wheel
- Few best practices, wasted dev cycles.
- Multiple, circuitous routes to the same place
- Can we be pioneers without so much chaos?

We’ve created a monster!

Without standards Gen AI RAG can be a beast!

The solution

A platform that tames the chaos

Open source platform that organizes GEN AI chaos

THE LINUX FOUNDATION

Open Platform for Enterprise AI

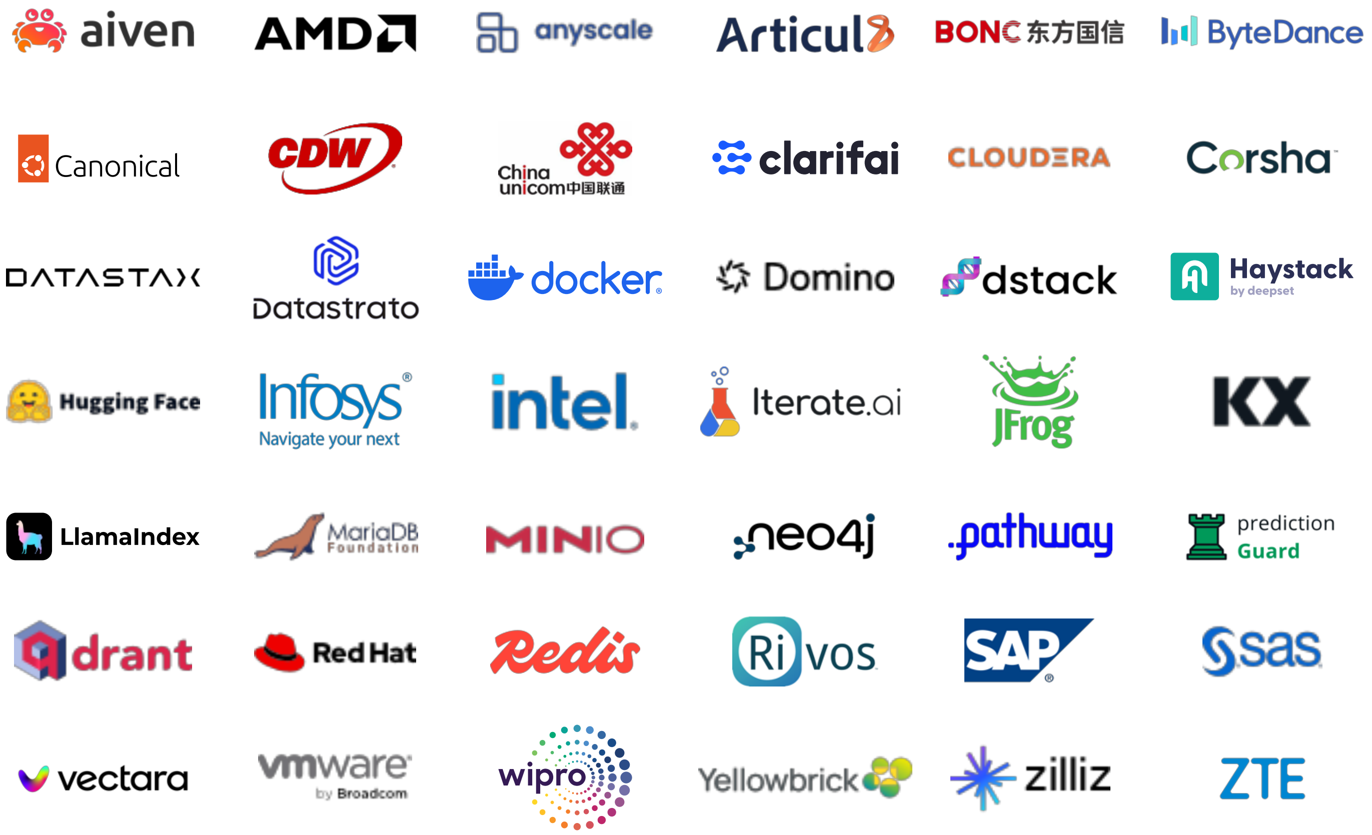
- Composable building blocks for generative AI systems
- Integrates LLMs, data stores, and prompt engines

  RAG AI blueprint with component stack structure and end-to-end workflow
- Additional stacks for translation, code generation, images
- Generative AI assessment tool for performance, features, trustworthiness and enterprise-grade readiness

OPEA Contributors

THE LINUX FOUNDATION

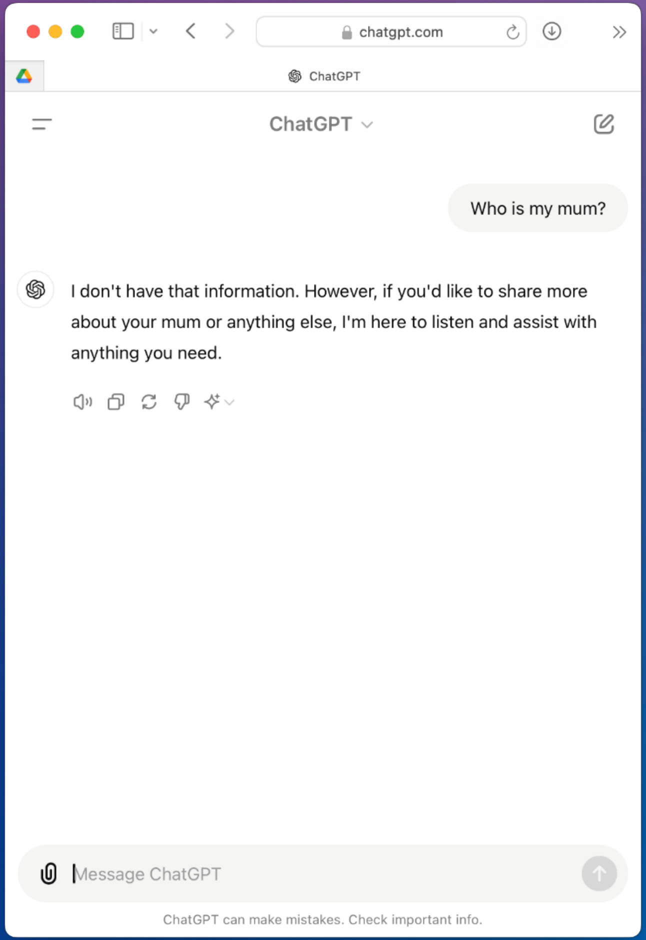
Open Platform for Enterprise AI



What is RAG?

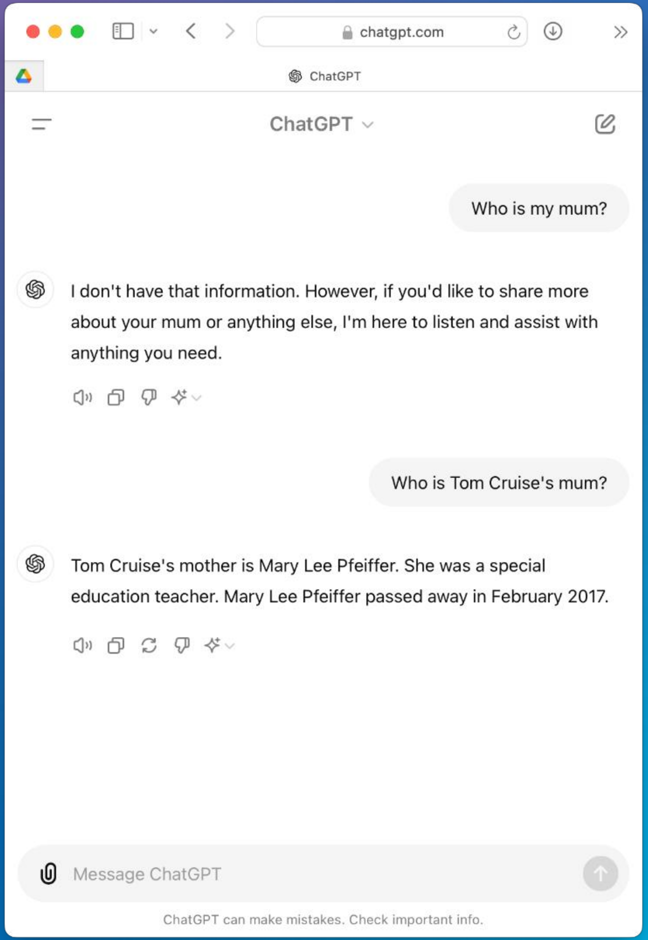
What is RAG

LLMs don’t know everything …



Who is my mum?

I don't have that information. However, if you'd like to share more about your mum or anything else, I'm here to listen and assist with anything you need.



Who is my mum?

I don't have that information. However, if you'd like to share more about your mum or anything else, I'm here to listen and assist with anything you need.

Who is Tom Cruise's mum?

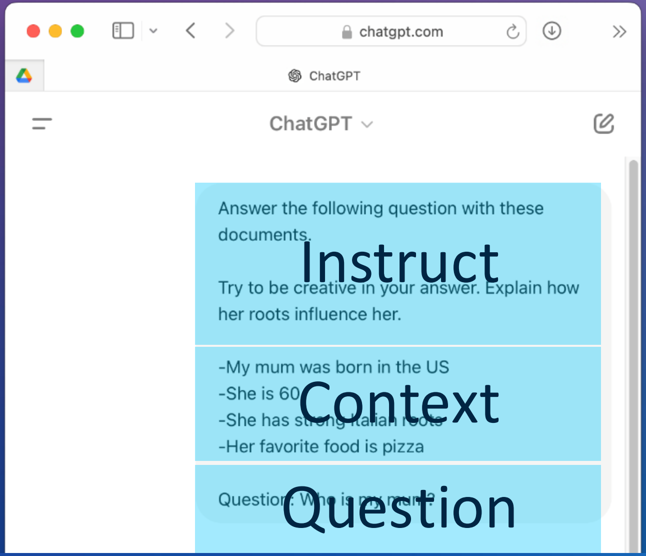
Tom Cruise's mother is Mary Lee Pfeiffer. She was a special education teacher. Mary Lee Pfeiffer passed away in February 2017.

But they do know a lot about their training data

YOU TOLD ME YOU’RE SMART!
I’m a pre-trained model with general knowledge. You should have given me context.

What is RAG

If we augment an LLM with new context …



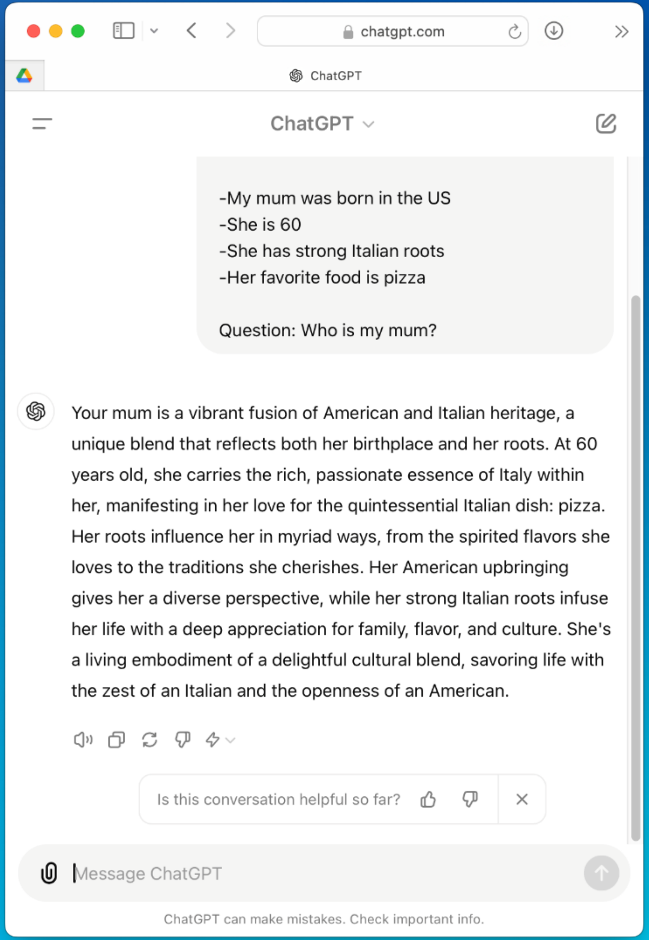
Answer the following question with these documents.

Try to be creative in your answer. Explain how her roots influence her.

-My mum was born in the US
-She is 60
-She has strong Italian roots
-Her favorite food is pizza

Question: Who is my mum?

It can generate a response to a novel topic



chatgpt.com
ChatGPT
ChatGPT
Is this conversation helpful so far?
Message ChatGPT
ChatGPT can make mistakes. Check important info.

-My mum was born in the US
-She is 60
-She has strong Italian roots
-Her favorite food is pizza

Question: Who is my mum?

Your mum is a vibrant fusion of American and Italian heritage, a unique blend that reflects both her birthplace and her roots. At 60 years old, she carries the rich, passionate essence of Italy within her, manifesting in her love for the quintessential Italian dish: pizza. Her roots influence her in myriad ways, from the spirited flavors she loves to the traditions she cherishes. Her American upbringing gives her a diverse perspective, while her strong Italian roots infuse her life with a deep appreciation for family, flavor, and culture. She's a living embodiment of a delightful cultural blend, savoring life with the zest of an Italian and the openness of an American.

RETRIEVAL

🧠

Retrieve data from a knowledge base

(my brain in this case!)

AUGMENTED

✨

Give the model instructions

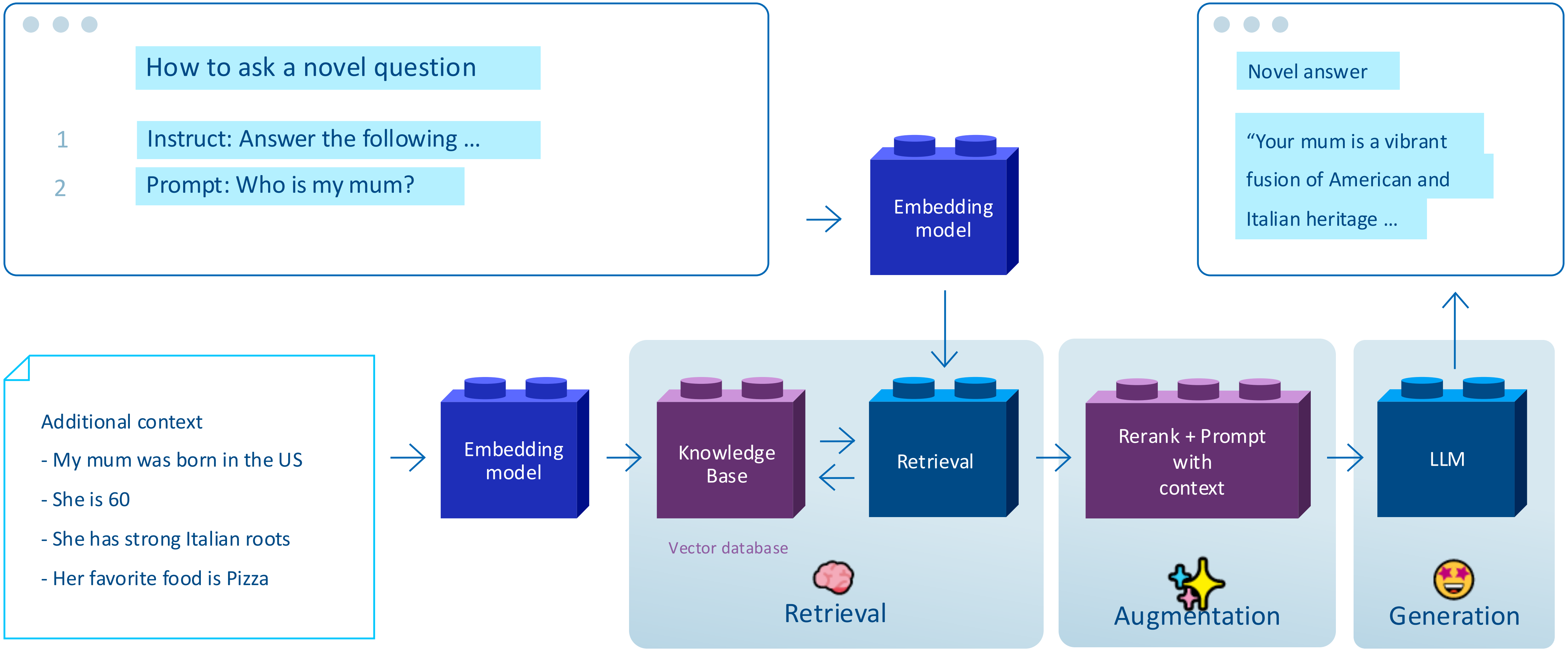
and context with the prompt

GENERATION

🤩

Get a response based on the retrieved data

RAG process flow



How to ask a novel question

1 Instruct: Answer the following …

2 Prompt: Who is my mum?

Additional context

- My mum was born in the US
- She is 60
- She has strong Italian roots
- Her favorite food is Pizza

Novel answer

“Your mum is a vibrant fusion of American and Italian heritage …

Lack of standards make Gen AI extremely complex

Ecosystem complexity for Gen AI

Data

- Indexing the data
- Creating embeddings (vector representations)
- Choosing storage (ISVs?)
- Preparing the data

LLM

- Selecting the right model
- Interacting with multiple models
- Choosing frameworks

Deploy

- Hardware efficiency
- Retrieval method
- Applying guardrails
- Retrieval accuratcy (reranking)
- Putting it all together
- Finding/building a megaservice

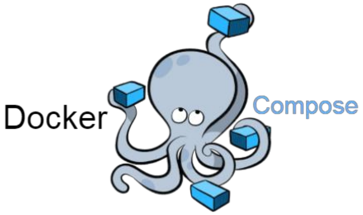
Open Platform for Enterprise AI (OPEA)

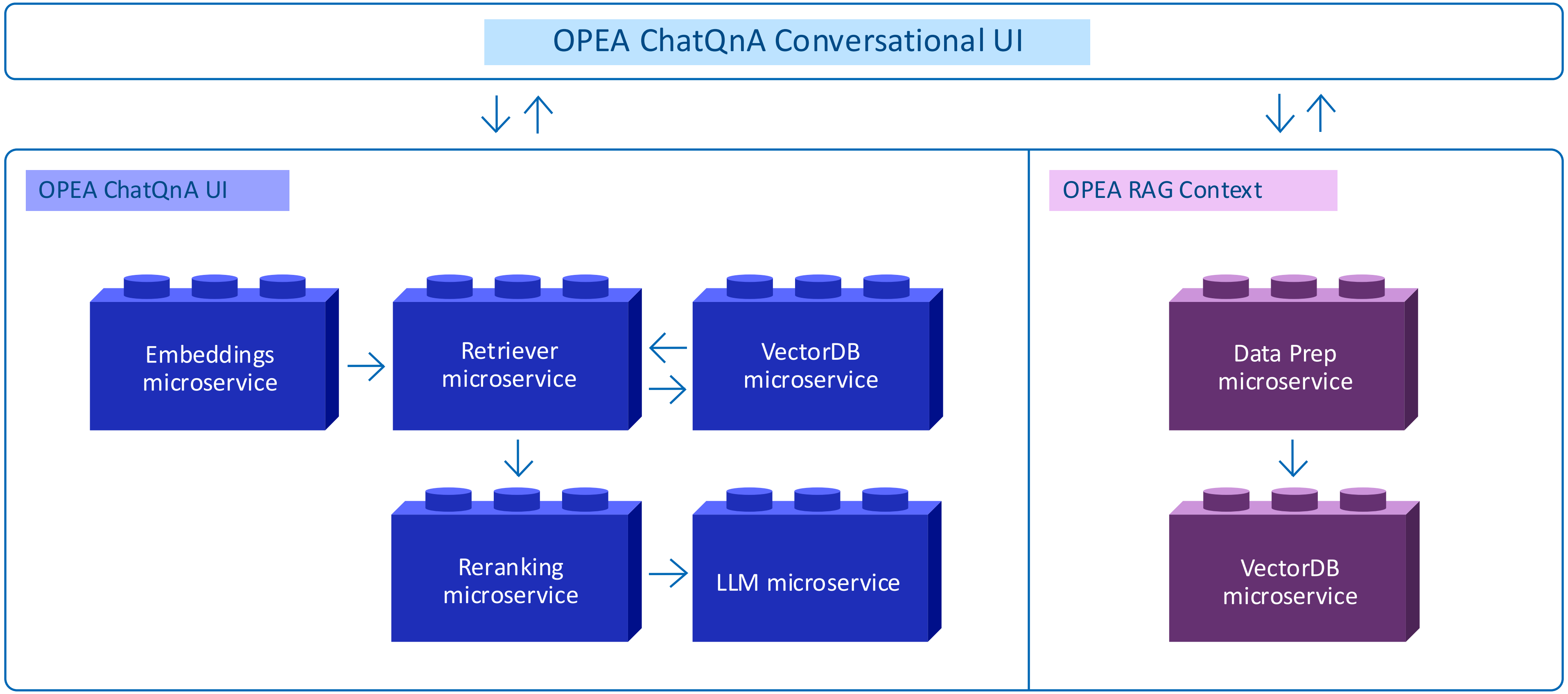
OPEA RAG reference meets these challenges

- Composable microservices for data, LLMs, and deployment
- Plug and play with 3rd party services—point to whichever data providers, embedding engines, and models you choose
- Over-arching mega service in place for faster development

DEMO

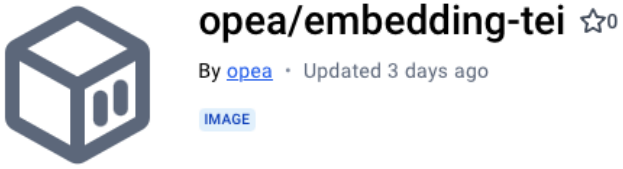
Under the hood – OPEA ChatQnA with RAG





- Chat with RAG running on microservice architecture
- Microservices can point to choice of services/APIs

Embeddings microservice - Converts from text to embeddings (vectors)



opea/embedding-tei ☆0

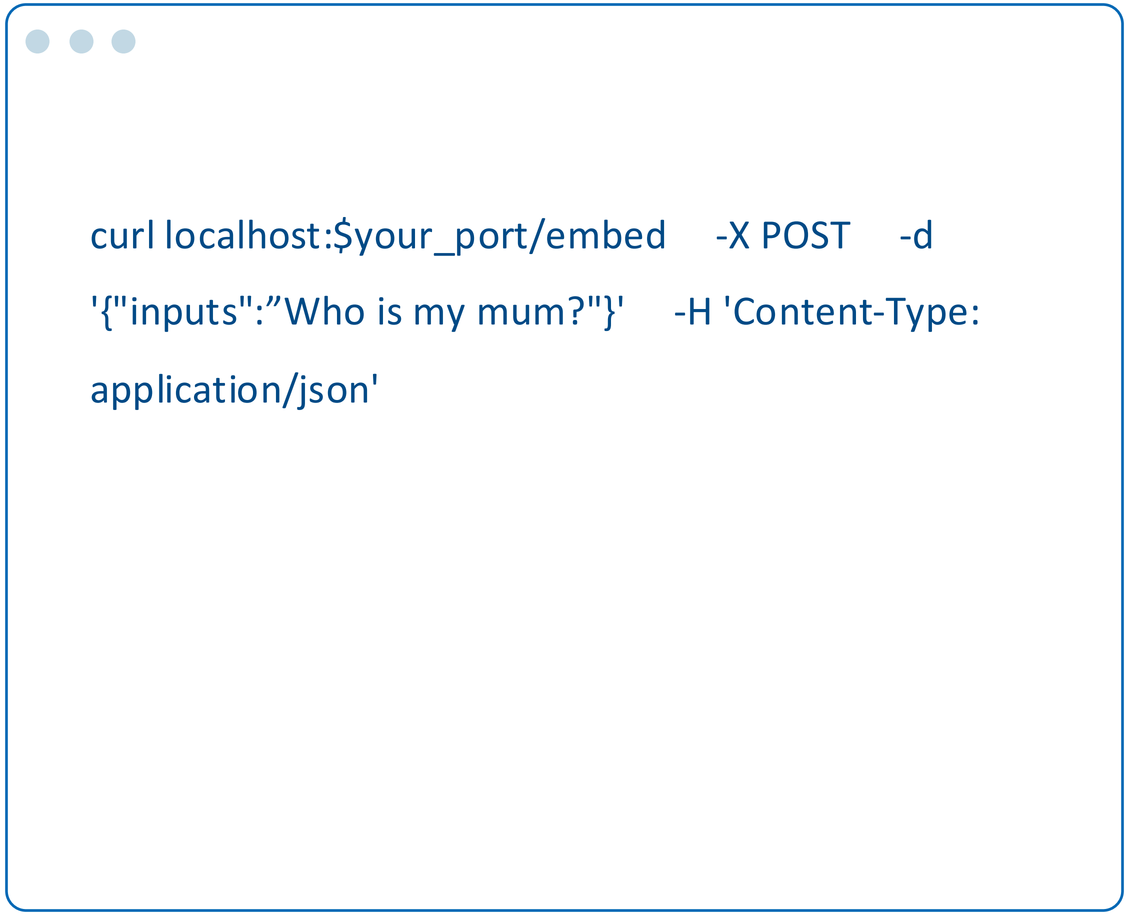
By opea · Updated 3 days ago

IMAGE

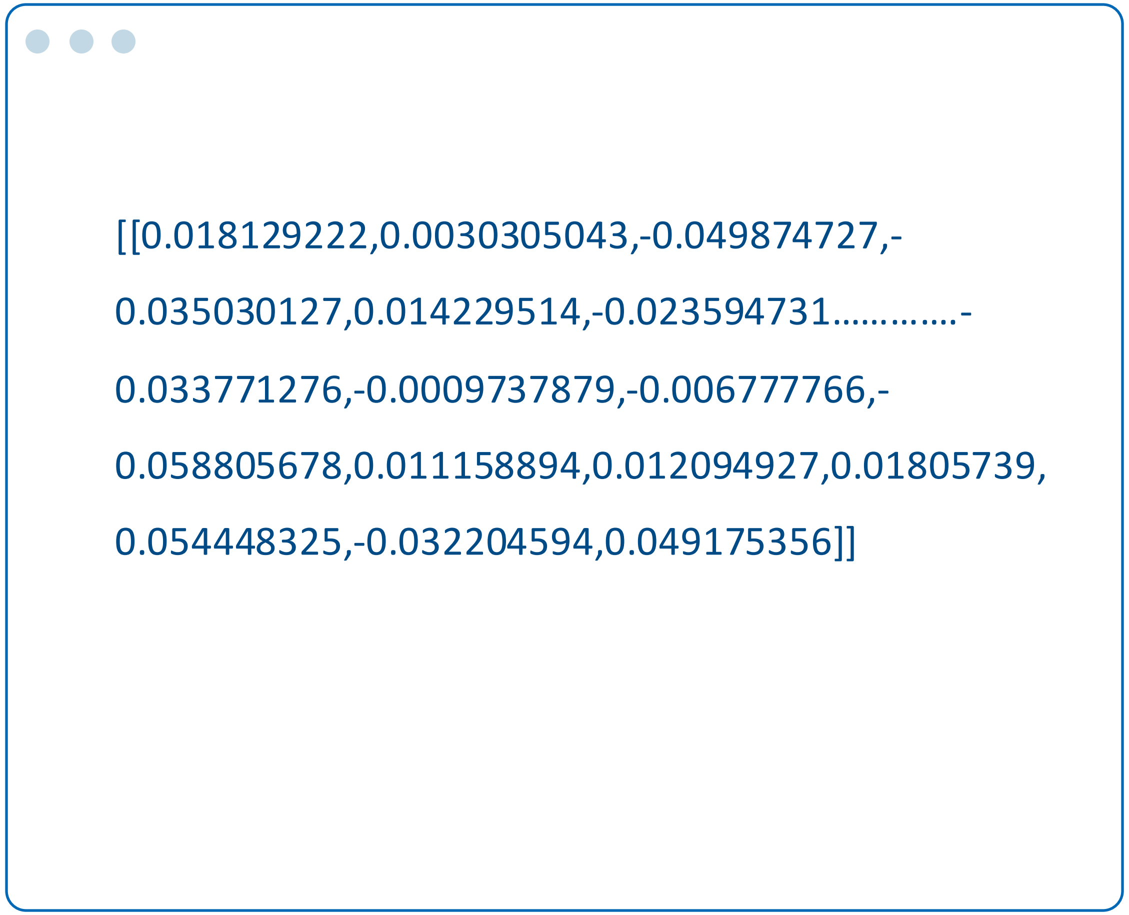
Variables

```
services:
  embedding:
    image: opea/embedding-tei:latest
    container_name: embedding-tei-server
    ports:
      - "6000:6000"
    ipc: host
    environment:
      http_proxy: ${http_proxy}
      https_proxy: ${https_proxy}
      TEI_EMBEDDING_ENDPOINT: ${TEI_EMBEDDING_ENDPOINT}
      LANGCHAIN_API_KEY: ${LANGCHAIN_API_KEY}
    restart: unless-stopped
```

Vector for “Who is my mum?”

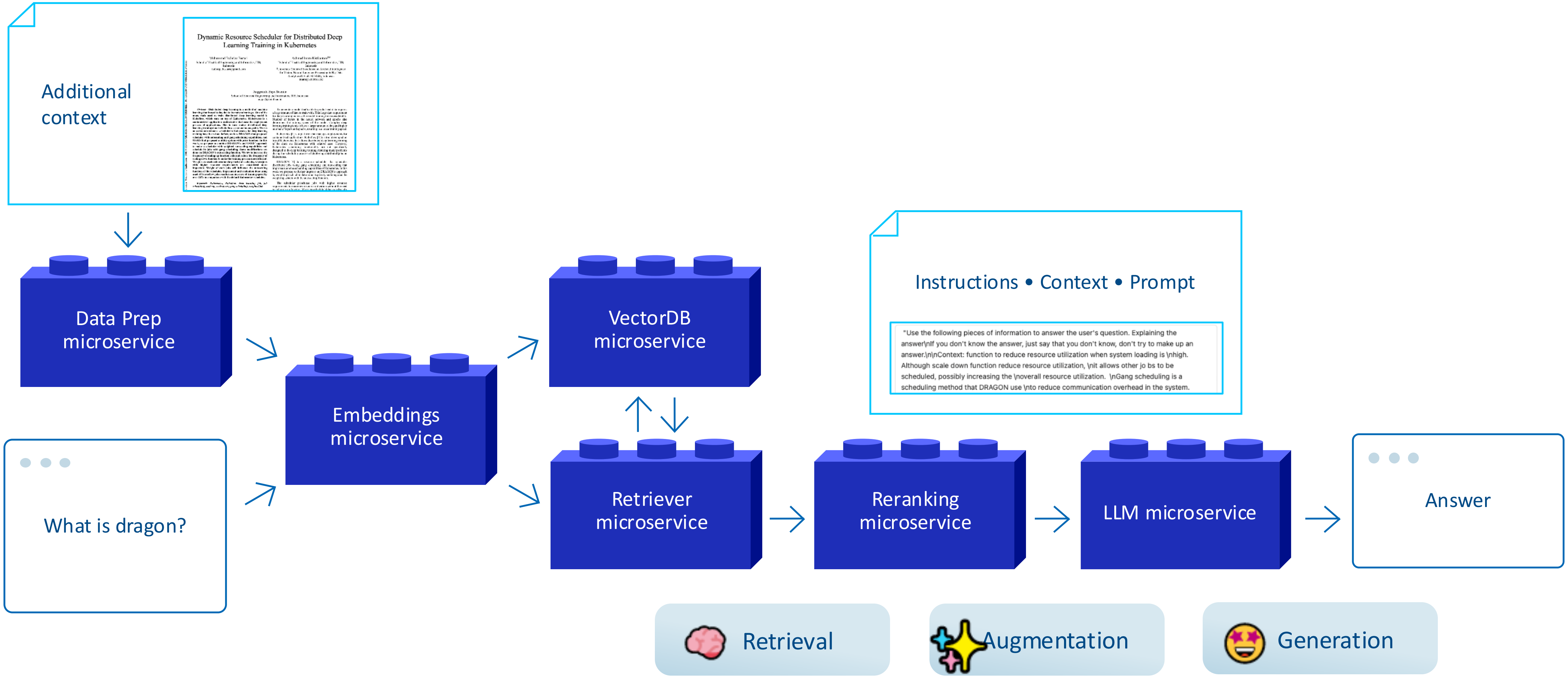


```
curl localhost:$your_port/embed    -X POST    -d
'{"inputs":”Who is my mum?"}'    -H 'Content-Type:
application/json'
```



```
[[0.018129222,0.0030305043,-0.049874727,-
0.035030127,0.014229514,-0.023594731………….-
0.033771276,-0.0009737879,-0.006777766,-
0.058805678,0.011158894,0.012094927,0.01805739,
0.054448325,-0.032204594,0.049175356]]
```

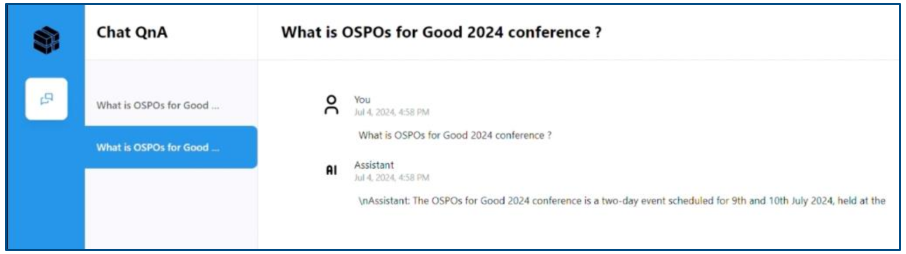
How OPEA ChatQnA answers on context



Dynamic Resource Scheduler for Distributed Deep Learning Training in Kubernetes

"Use the following pieces of information to answer the user's question. Explaining the answer\nIf you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nContext: function to reduce resource utilization when system loading is \nhigh. Although scale down function reduce resource utilization, \nit allows other jo bs to be scheduled, possibly increasing the \noverall resource utilization. \nGang scheduling is a scheduling method that DRAGON use \nto reduce communication overhead in the system.

OPEA ChatQnA Megaservice



Chat QnA

What is OSPOs for Good ...

What is OSPOs for Good ...

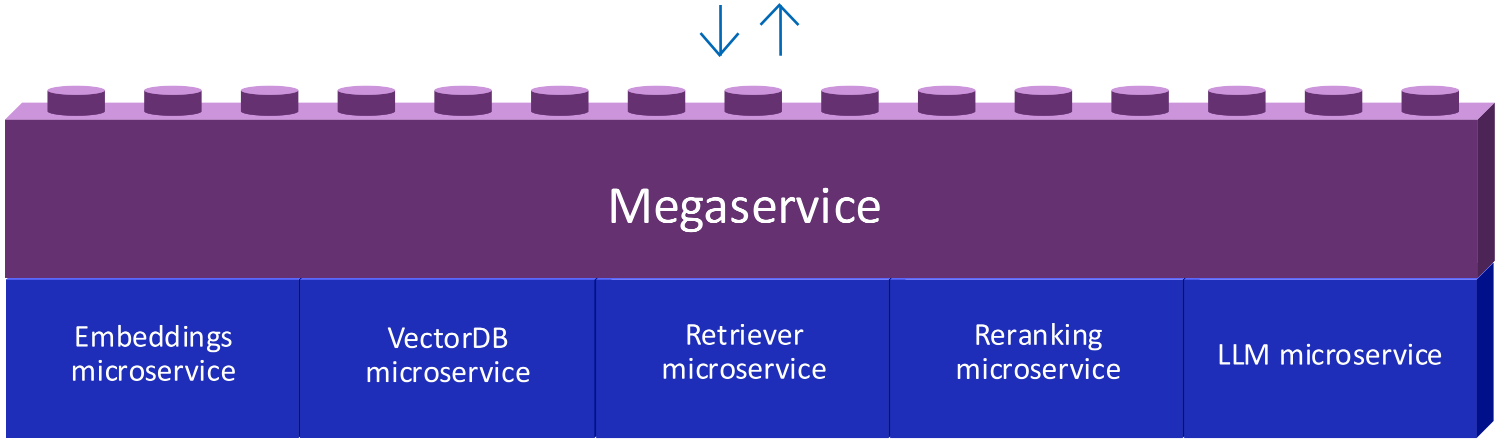
What is OSPOs for Good 2024 conference ?

You
Jul 4, 2024, 4:58 PM

What is OSPOs for Good 2024 conference ?

AI Assistant
Jul 4, 2024, 4:58 PM

\nAssistant: The OSPOs for Good 2024 conference is a two-day event scheduled for 9th and 10th July 2024, held at the



Megaservice

Embeddings microservice

VectorDB microservice

Retriever microservice

Reranking microservice

LLM microservice

CODE

Check out ChatQnA on GitHub

RAG conclusions and challenges

The good

- RAG is a great approach to having context-based answers
- OPEA makes the building blocks easy to deploy

The challenges

- Dealing with a large amount of data (Play with different similar search techniques)
- Different types of data (Images/Tables) – Intel Tech on Medium articles
- Is it all about indexing? Explore Retrieval augmented text-to-SQL generation

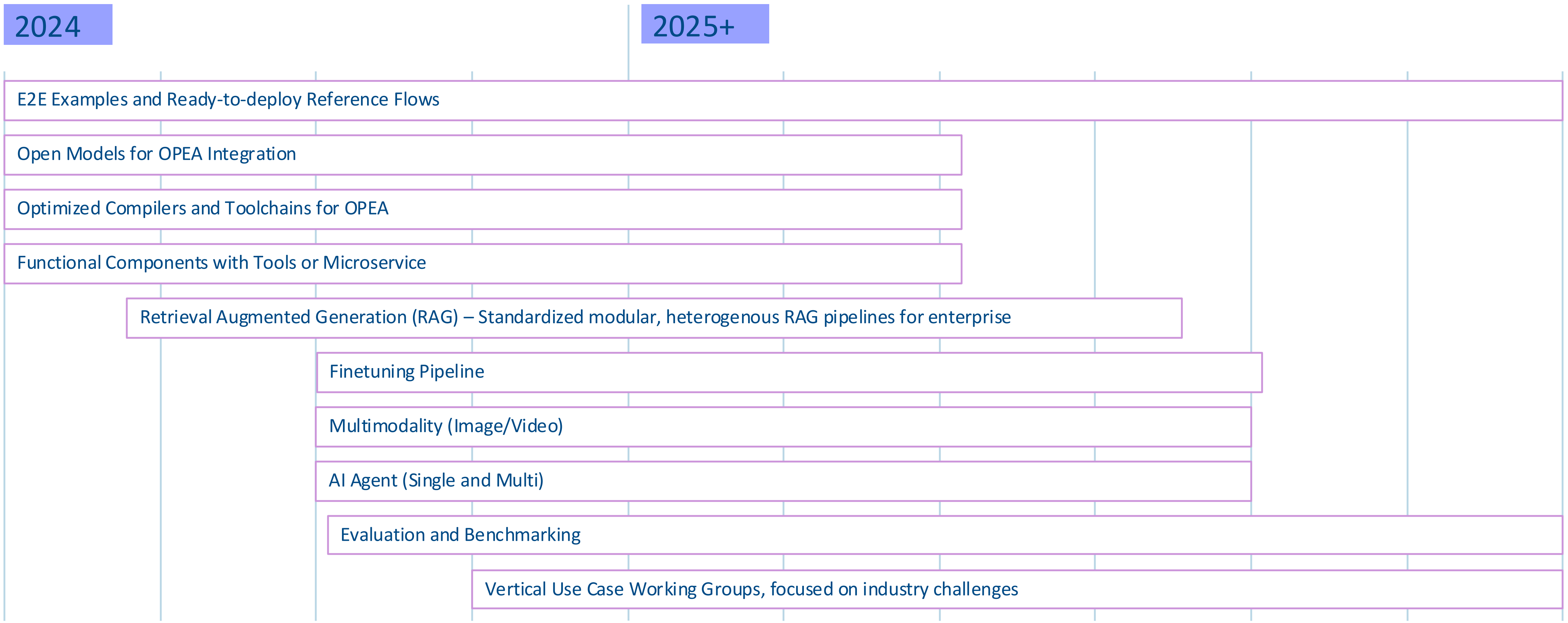
THE LINUX FOUNDATION

Open Platform for Enterprise AI

Join us!

There is more—a lot more—you can build!

OPEA* Roadmap



2024

2025+

E2E Examples and Ready-to-deploy Reference Flows

Open Models for OPEA Integration

Optimized Compilers and Toolchains for OPEA

Functional Components with Tools or Microservice

Retrieval Augmented Generation (RAG) – Standardized modular, heterogenous RAG pipelines for enterprise

Finetuning Pipeline

Multimodality (Image/Video)

AI Agent (Single and Multi)

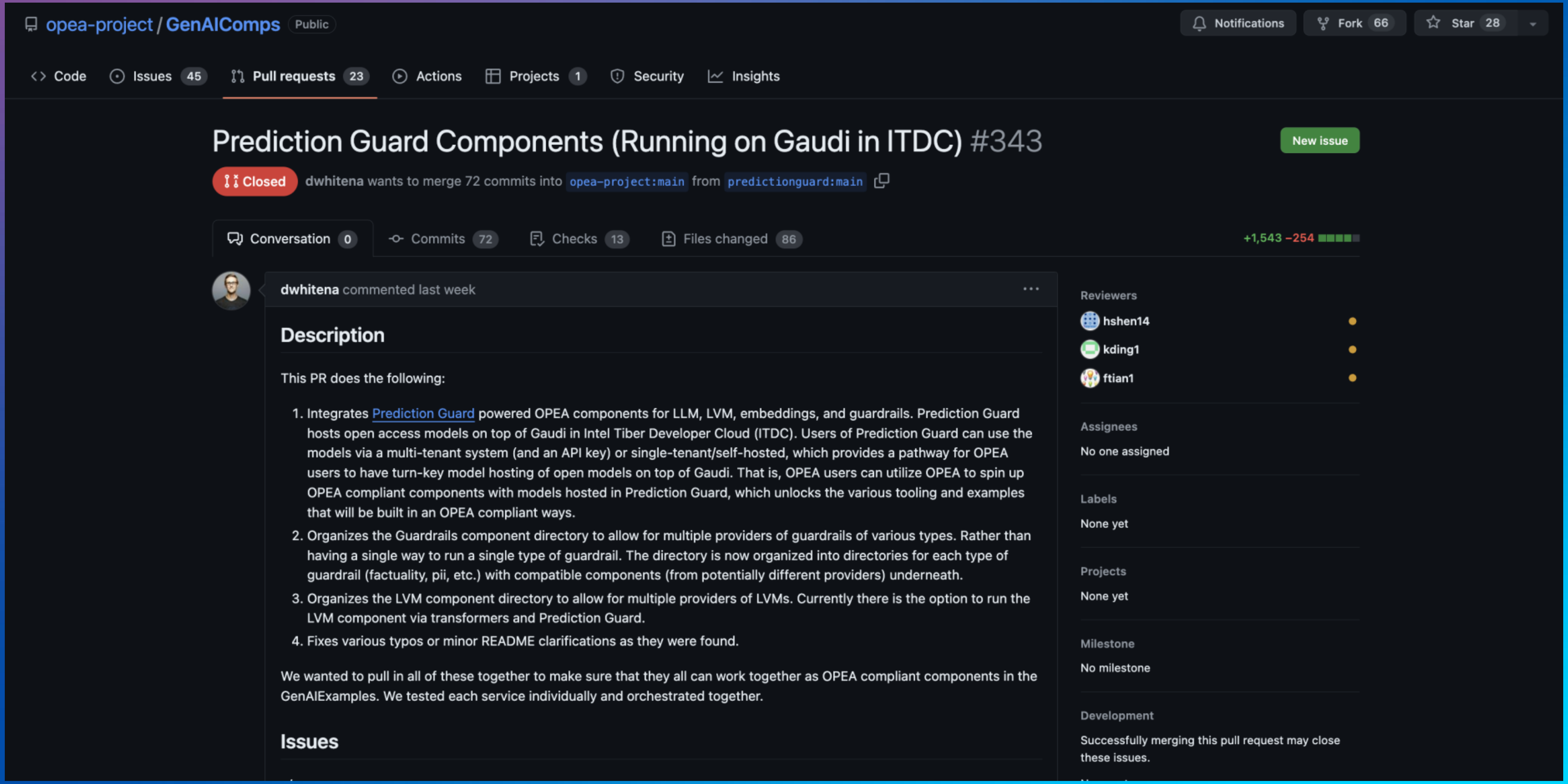
Evaluation and Benchmarking

Vertical Use Case Working Groups, focused on industry challenges

*Preliminary

Open Platform for Enterprise AI

Help build a better OPEA, contribute!



Prediction Guard Components (Running on Gaudi in ITDC) #343

Closed dwhitena wants to merge 72 commits into `opea-project:main` from `predictionguard:main`

dwhitena commented last week

Description

This PR does the following:

1. Integrates Prediction Guard powered OPEA components for LLM, LVM, embeddings, and guardrails. Prediction Guard hosts open access models on top of Gaudi in Intel Tiber Developer Cloud (ITDC). Users of Prediction Guard can use the models via a multi-tenant system (and an API key) or single-tenant/self-hosted, which provides a pathway for OPEA users to have turn-key model hosting of open models on top of Gaudi. That is, OPEA users can utilize OPEA to spin up OPEA compliant components with models hosted in Prediction Guard, which unlocks the various tooling and examples that will be built in an OPEA compliant ways.
2. Organizes the Guardrails component directory to allow for multiple providers of guardrails of various types. Rather than having a single way to run a single type of guardrail. The directory is now organized into directories for each type of guardrail (factuality, pii, etc.) with compatible components (from potentially different providers) underneath.
3. Organizes the LVM component directory to allow for multiple providers of LVMs. Currently there is the option to run the LVM component via transformers and Prediction Guard.
4. Fixes various typos or minor README clarifications as they were found.

We wanted to pull in all of these together to make sure that they all can work together as OPEA compliant components in the GenAIExamples. We tested each service individually and orchestrated together.

Issues

Reviewers

- hshen14
- kding1
- ftian1

Assignees

No one assigned

Labels

None yet

Projects

None yet

Milestone

No milestone

Development

Successfully merging this pull request may close these issues.

Participate in OPEA

- Join a Working Group
- Bring Enterprise AI use cases
- Contribute code, docs, projects, blueprints, and more
- Provide feedback
- Evangelism and Promotion of OPEA in YOUR communities, events

Visit OPEA.dev

Scan for tools, sources, and resources

Visit the Intel Open Ecosystem Community and Evangelism GitHub page

- A PDF of this presentation
- Guides and community resources
- Links to articles and source material

Where to find Katherine

- LinkedIn: /katherinedruckman
- Fediverse: @katherined@reality2.social
- Twitter/X: @katherined

Where to find Ezequiel

- LinkedIn: /ezelanza
- Fediverse: @eze_Lanza@mstdn.ca
- Twitter/X: @eze_lanza

Open Platform for Enterprise AI

Thank you!

intel®

Notices and disclaimers

Intel is committed to respecting human rights and avoiding complicity in human rights abuses. See Intel's Global Human Rights Principles. Intel® products and software are intended only to be used in applications that do not cause or contribute to a violation of an internationally recognized human right.

Intel® technologies may require enabled hardware, software, or service activation. No product or component can be absolutely secure. Your costs and results may vary.